Second International Conference
Modelling and Development of Intelligent Systems
Sibiu - Romania, September 29 - October 02, 2011

Estimation of the selectivity factor for a set of queries

Letiţia Velcescu, Laurenţiu Vasile

Abstract

In this paper, we study the selectivity factor, extending this concept to the case of a queries set $Q_1, ..., Q_n$. These queries are performed on the tables of a transactional database. So, they are supposed to be updated dynamically and, consequently, the selectivity factor associated to each query may vary in different moments. Because the selectivity factor has an important role in database optimization, it is necessary to be able to estimate it. We propose an algorithm for the estimation of the generalized selectivity factor, the concept we introduce, and also a hybrid estimator for it.

1 Introduction

Because the volume of information is growing continuously and fastly, the information processing and the quick access to it remain important issues in the information technology. In this paper, we present some factors which influence the execution of a query in a database management system (DBMS), such as the storage and access methods, and the selectivity factor of the queries.

The DBMSs include optimizers, which are software modules responsible with the adequate choice of a query's execution plan. Also, the DBMSs contain tools which allow the database administrator to tune the database such that it works appropriately to the specific of the applications. A more refined concept which appeared in this context is the one of database auto-administration (Aouiche et al 2003). This concept supposes the existence of some modules which build up the structures necessary to the optimum execution of an application.

Generally, the optimizer chooses the execution plan according to the query type and the available statistics generated by the system.

The optimization of the SQL code is also important because it has a significant impact over the resources and, consequently, on the database server's performances (Connolly, Begg, 2002).

The selectivity factor is presented in the second part of the paper. The third part is dedicated to the generalization of this concept, associated to one query so far, to the one associated to a query set. This part includes the estimation problems we deal with in this context.

2 The selectivity factor

A query's selectivity factor is extremely useful in the performance evaluation studies and in the record linkage problems. So, it is interesting that this factor could be estimated. In the recent research, the stochastic modelling was used in order to estimate this factor in the databases with uncertain information. The selectivity concept was used in the analysis of the performances, of the files organization, in the physical design of the databases (selection of indexes, partitions or attribute groups), as well as in the queries optimization. From this latter point of view, the definition of a statistic profile was taken in consideration, including the information stored or estimated in the database in order to optimize it. Such a statistic profile may include the number of tuples in a relation, the number of distinct values in each domain, the average number of records in a data block etc. Based on these statistics, the optimizor should determine: the cost of the individual operations specific to the relational databases, the new statistic profiles of the relations derived from operations, the cost of a sequence of operations.

Definition 1 (Delobel, Adiba, 1982) *Given a relation R, the selectivity factor of a condition C, denoted by $Sf(C)$, is the probability that a record in R satisfies the condition C.*

3 The generalization of the selectivity factor for a queries set

In this part, we propose to extend the concept of selectivity factor associated to a query, to the notion of selectivity factor associated to a queries set $Q_1, ..., Q_l$. The framework in which we will place this concept supposes that the tables of the database are updated dynamically. So, at distinct moments of time, the selectivity factor associated to each query is different.

Suppose that the query Q_i has the selectivity factor $p_i = \frac{\alpha_i}{k}$, where k is the number (constant) of lines of the table, and α_i is the number of lines selected by the query Q_i. In this case, the selectivity factor of the queries set can be estimated by the variable:

$$\bar{p} = \frac{1}{l}\sum_i p_i = \frac{1}{l}\sum_i \frac{\alpha_i}{k} = \frac{1}{lk}\sum_i \alpha_i \tag{1}$$

The mean of this variable is $E(\bar{p}) = p$ and it represents the (theoretical) selectivity factor of the queries set, where $\bar{p}$ is an unbiased estimation for p. The dispersion of this variable is $\frac{p(1-p)}{l} \leqslant \frac{1}{4l}$, and the standard deviation is $\sigma \leqslant \dfrac{1}{2\sqrt{l}}$.

From the Chebyshev's inequality, we get:

$$P(|\bar{p} - p| < t\sigma) \geq 1 - \frac{1}{t^2} \tag{2}$$

We denote $1 - \frac{1}{t^2} = 1 - \delta$, and from here it results that $t_\delta = \dfrac{1}{\sqrt{\delta}}$.

We denote $t\sigma = \varepsilon \simeq 0.01$. For such a value ($\varepsilon$ small), we obtain a big value $1 - \delta$, so δ is small.

From the preceding notations, it implies that:

$$t_\delta \cdot \frac{1}{2\sqrt{l}} \leq 0.01 \tag{3}$$

Therefore, in order to be able to estimate the selectivity factor in the way described previously, the number l of queries must satisfy the condition $l \geq \dfrac{t_\delta^2}{4 \cdot (0.01)^2}$.

Supposing that the tables are updated dynamically, we consider that the number of lines of the tables at the moments of different queries is k_i, uniformily distributed. We suppose that the minimum, respectively maximum values of this variable are known: $m \leq k_i \leq s$. Then, by means of the simulation methods of the uniform variable (Vaduva, 2004), we obtain an estimation of the generalized selectivity factor for a queries set through the algorithm AEGSF presented below (Algorithm 1).

On the basis of the values $p_1, ..., p_N$, generated at the step 6 of this algorithm, the selectivity factor of the queries set can be estimated by $\bar{p}$.

Algorithm 1 AEGSF (Algorithm for the estimation of the generalized selectivity factor)

```
Require: The value of $t_\delta$.
Ensure: The value of $\bar{p}$.
$N_0 \leftarrow \left[\frac{t_\delta^2}{4 \cdot (0.01)^2}\right] + 1$
Choose $N \geq N_0$
for $i = 1$ TO $N$ do
    Generate a random value $U \in (0, 1)$
    Determine $k_i = m + [(s - m) \cdot U] + 1$
    Calculate $p_i = \frac{\alpha_i}{k_i}$
end for
return $\bar{p} = \frac{1}{N} \sum_i p_i$
```

3.1 Estimation of the selectivity factor

In many cases, in the frame of the statistic profiles used in the previous research, it was supposed that the tuples of the relations are identically distributed, relative to the values of an attribute, and the values of different attributes are independent. This supposition is not realistic, in many environments in which databases have an important role, because the values of the attributes can be imprecise and there can be dependencies between attributes. Because of this reason, the preoccupations concerning the estimation of this factor concerned parametric, non-parametric methods, and also the maximum entropy principle (Christodoulakis, 1989).

A subsequent classification of the methods of estimation of the selectivity factor used traditionally concerned the following estimator types (Ling, Sun, 1999):

- based on selection, which determine the selectivity factor only on the basis of the information at runtime, without using the information collected previously;
- parametric and table based, which use only the information collected previously, ignoring the on-line information.

The disadvantage of the first class of estimators consists in the insufficient use of the available information, whereas the second class leads to an imprecise estimation in an environment with frequent updates. There have been proposed hybrid estimators, which weight the two types enumerated above and whose results have been validated in practice.

3.2 Hybrid estimator for the generalized selectivity factor

We will present a hybrid estimator (Ling, Sun, 1999) for the selectivity factor of a query and we will extend the results for the case of the selectivity factor associated to a queries set, introduced before.

In the case of a single query, we consider f the characteristic function of a selection predicate, and be x_i a tuple. The function f can be defined in the following way:

$$y_i = f(x_i) = \begin{cases} 1, \text{ if } x_i \text{ satisfies the selection predicate} \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

In the approach based on selection, the main technique consists in choosing randomly, repeatedly, a tuple in the table on the basis of the query's predicate, followed by the realization of the inference about the real selectivity, using the estimated selectivity obtained from the sample data. Thus, one can realize the inference according to which an approximation of the real selectivity p is:

$$\hat{p}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} y_{r_i} = \frac{1}{n} \cdot \sum_{i=1}^{n} f(x_{r_i}) \tag{5}$$

where n represents the sample size, k is the total number of tuples and the index r_i is a random integer, between 1 and k. Consequently, the average total number of tuples which satisfy the selection is $k \cdot \hat{p}_n$.

A hybrid estimator $\breve{p}_n$ of the selectivity factor is given by a linear combination between the estimated selectivity $\hat{p}_n$ and the estimated selectivity $\tilde{p}$, obtained by a parametric estimator or by a table based estimator:

$$\breve{p}_n = t \cdot \hat{p}_n + (1-t) \cdot \tilde{p} \tag{6}$$

where t is a parameter in the interval $[0,1]$.

In order to validate an estimator, the mean-squared error (*mse*) is used to quantify the estimators performances:

$$mse = E(\bar{p}-p)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\bar{p}_i - p)^2 \tag{7}$$

where $\bar{p}_i$ is the individual selectivity estimated by an estimator, p is the total, real selectivity, depending on the given query, and n is the sample size.

The value *mse* of an estimator represents the accuracy of its estimation, as well as its safety. The smaller the *mse* value, the better the estimator is. For a method based on selection, the *mse* value is $\frac{p \cdot (1-p)}{n}$. For an estimator which uses a parametric or table based method, the *mse* value is $(\tilde{p}-p)^2$, and $\tilde{p}$ remains unchanged for the given query, until the parametric or table based estimator is recomputed using the updated information.

The different values of the parameter t represent different weights of the two estimators. In the extreme cases $t = 1$ or $t = 0$, the hybrid model reduces to a selection based estimator, respectively to a parametric or a table based one. The existence of an optimum value of the parameter t and the calculation of this value have been determined in the following theorem.

Theorem 1 (Ling and Sun, 1999) *The optimum value of t, denoted by t_n^*, is given by the formula:*

$$t_n^* = \frac{(\tilde{p}-p)^2}{p \cdot (1-p)/n + (\tilde{p}-p)^2} \tag{8}$$

The mse value for the hybrid estimator corresponding to the optimum parameter t_n^* is smaller than each of the two estimators when $0 < p < 1$ and $p \neq \tilde{p}$, meaning that:

$$E(\breve{p}_n^* - p)^2 < \min\{\frac{p \cdot (1-p)}{n}, (\tilde{p}-p)^2\} \tag{9}$$

where $\breve{p}_n^*$ is $\breve{p}_n$ for $t = t_n^*$.

We can apply the hybrid estimators in the case proposed previously, of several queries, therefore of the generalized selectivity factor. We know that $p_1, ...p_q$ are the selectivity factors associated to the q queries. Consider f_i the characteristic functions of the predicates associated to each of the q queries and be x_{ij} a tuple in the query i. We consider a selection associated to each query.

The functions f_i are given by the formula:

$$y_{ij} = f_i(x_{ij}) = \begin{cases} 1, \text{ if } x_{ij} \text{ satisfies the selection predicate} \\ \quad \text{of the query } i \\ 0, \text{ otherwise} \end{cases} \tag{10}$$

Be n_i the sizes of the samples associated to the q queries. Then, an approximation of the selectivity of the query i is given by:

$$\hat{p}_{n_i} = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} y_{ir_j} = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} f_i(x_{ir_j}) \tag{11}$$

where r_j is a random integer, between 1 and n_i.

The approximate number of tuples of the result of the query i will be $k_i \cdot \hat{p}_{n_i}$, where k_i is the total number of tuples of the relation in the query i.

Every query has an associated hybrid estimator:

$$\breve{p}_{n_i} = t_i \cdot \hat{p}_{n_i} + (1 - t_i) \cdot \tilde{p}_i \tag{12}$$

For the queries set, we will have:

$$\bar{p} = \frac{1}{N} \cdot \sum_i \breve{p}_{n_i} \tag{13}$$

where N was determined previously.

Algorithm 2 AEGSFHE (Algorithm for the estimation of the generalized selectivity, using hybrid estimators)

```
Require: The value of $t_\delta$.
Ensure: The value of $\bar{p}$.
1: $N_0 \leftarrow \left[\frac{t_\delta^2}{4 \cdot (0.01)^2}\right] + 1$
2: Choose $N \geq N_0$
3: for $i = 1$ TO $N$ do
4:     Generate a random value $U \in (0, 1)$
5:     Determine $n_i = m + [(n - m) \cdot U] + 1$
6:     Calculate $\breve{p}_{n_i} = t_i \cdot \hat{p}_{n_i} + (1 - t_i) \cdot \tilde{p}_i$
7: end for
8: return $\bar{p} = \frac{1}{N} \sum_i \breve{p}_{n_i}$
```

Thus, the algorithm AEGSF becomes AEGSFHE.

On the basis of the values $\breve{p}_{n_1}, ..., \breve{p}_{n_N}$ generated at step 6 of the algorithm, the selectivity factor of the queries set can be estimated by $\bar{p}$ which is computed in the last step of the algorithm.

4 Conclusions

This paper refered some problems related to the query optimization in database management systems. We mentioned some factors which interfere in this process, such as some physical parameters of the system on the access to data in the databases. We studied the queries' selectivity factor, which has an important role in database optimization. We proposed the generalization of this concept for a set of queries. The framework in which this concept becomes useful is the one in which the tables of the database update dynamically, so that the selectivity associated to a query may vary. We proposed an algorithm for estimation of the generalized selectivity factor, which uses simulation methods of the uniform univariate.

The methods for the estimation of the selectivity factor proposed so far concerned sample based estimators, parametric and table based estimators. Each of these classes present some disadvantages, reason for which the research in this field proposed the introduction of hybrid estimators, which weight these types and whose results have been validated in practice.

For the concept of generalized selectivity factor that we introduced, we also proposed a hybrid estimator formalized as an algorithm.

Acknowledgement: This work was supported by the strategic grant POSDRU/89/1.5/S/58852, Project "Postdoctoral programme for training scientific researchers" cofinanced by the European Social Found within the Sectorial Operational Program Human Resources Development 2007-2013.

References

[1] Aouiche, K., Darmont, J., Gruenwald, L., *Frequent itemsets mining for database auto-administration*, In Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS03), 2003, p. 98-103.

[2] Christodoulakis, S., *On the estimation and use of selectivities in database performance evaluation*, Research Report CS 89-24, Department of Computer Science, University of Waterloo, Canada, 1989 (http://www.cs.uwaterloo.ca/research/tr/1989/CS-89-24.pdf).

[3] Connolly, T.M., Begg, C.E., *Database Systems: A Practical Approach to Design, Implementation and Management*, 3^{rd} edition, Addison-Wesley, 2002.

[4] Delobel, C., Adiba, M., *Bases de donnes et systmes relationnels*, Dunod, 1982.

[5] Feng, J., Qian, Q., Liao, Y., Li, G., Ta, N., *DMT: A Flexible and Versatile Selectivity Estimation Approach for Graph Query*, WAIM 2005, LNCS 3739, Springer, 2005, p. 663 669.

[6] Lim L., Wang M., Padmanabhan S., Vitter J., Parr R., *XPathLearner: An On-Line Self-Tuning Markov Histogram for XML Path Selectivity Estimation*, in Proceedings of the 28th VLDB Conference, 2002, p. 442-453.

[7] Ling, Y., Sun, W., *A Hybrid Estimator for Selectivity Estimation*, IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 2, IEEE Educational Activities Department, 1999, p. 338-354.

[8] Vaduva, I., *Modele de simulare*, Editura Universitatii din Bucuresti, 2004.

Letitia Velcescu
University of Bucharest
Faculty of Mathematics and Informatics
14 Academiei, 010014 Bucharest
ROMANIA
E-mail: *letitia@fmi.unibuc.ro*

Laurentiu Vasile
University of Bucharest
Faculty of Mathematics and Informatics
14 Academiei, 010014 Bucharest
ROMANIA
E-mail: *vsl@fmi.unibuc.ro*